

COMPONENT PART NOTICE

THIS PAPER IS A COMPONENT PART OF THE FOLLOWING COMPILATION REPORT:

TITLE: Workshop on Assessing Uncertainty Held in Monterey, California on November
13-14, 1986.

TO ORDER THE COMPLETE COMPILATION REPORT, USE AD-A180 850.

THE COMPONENT PART IS PROVIDED HERE TO ALLOW USERS ACCESS TO INDIVIDUALLY AUTHORED SECTIONS OF PROCEEDING, ANNALS, SYMPOSIA, ETC. HOWEVER, THE COMPONENT SHOULD BE CONSIDERED WITHIN THE CONTEXT OF THE OVERALL COMPILATION REPORT AND NOT AS A STAND-ALONE TECHNICAL REPORT.

THE FOLLOWING COMPONENT PART NUMBERS COMPRISE THE COMPILATION REPORT:

AD#: P005 291 thru P005 301 AD#: _____
AD#: _____ AD#: _____
AD#: _____ AD#: _____



**A BAYESIAN VIEW OF
ASSESSING UNCERTAINTY
AND COMPARING EXPERT
OPINION**

by

Morris H. DeGroot

Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Technical Report Number 387
January, 1987

Presented at the Workshop on Assessing Uncertainty, Naval Postgraduate School, Monterey, California, November 13-14, 1986. This research was supported in part by the National Science Foundation under grant DMS-8320618.

Abstract

A Bayesian approach to the problem of comparing experts or expert systems is presented. The question of who is an expert is considered and comparisons among well-calibrated experts are studied. The concept of refinement, in various equivalent forms, is used in this study. An informative example of the combination of the opinions of well-calibrated experts is described. Total orderings of the class of well-calibrated experts are derived from strictly proper scoring rules.

Keywords and phrases: Predictions; forecasters; well calibrated; expert systems; combining opinion; scoring rules.

1 Introduction

In the fields of artificial intelligence and expert systems, the necessity of assessing uncertainty and of coping with that uncertainty by developing methods for decision making under uncertainty are now widely recognized. In this paper, I will argue in favor of the Bayesian approach to assessing uncertainty, and then describe some ways in which this approach can be used to compare experts or expert systems.

The argument in favor of the Bayesian approach proceeds in two steps: (1) The quantitative assessment of uncertainty is in itself a sterile exercise unless that assessment is to be used to make decisions. (2) The Bayesian approach provides the only coherent methodology for decision making under uncertainty (see, e.g., Savage, 1954; DeGroot, 1970; or Lindley, 1987).

The Bayesian approach to the assessment of uncertainty is defined to be the approach in which any uncertainty about the values of various quantities on the part of the decision maker or the person receiving information from an expert or an expert system is represented by the person's subjective joint probability distribution for those values. Indeed, in the fields of artificial intelligence and expert systems, the terms "Bayesian approach" and "probability approach" are often used interchangeably. This usage is appropriate because the Bayesian approach is not characterized, as is sometimes stated, by the repeated use of Bayes' Theorem.

but by the ubiquitous specification of probabilities to represent uncertainty.

Two other approaches to the representation of uncertainty in expert systems that have been widely discussed are belief functions (Shafer, 1976, 1982, 1987) and fuzzy logic (Zadeh, 1979, 1983). Both of these approaches can provide reasonable approximations to probability under special conditions when it is not necessary for a decision maker to specify a fully-detailed, high-dimensional joint probability distribution for all of the quantities about which he or she is uncertain in order to be able to choose an effective decision. In general, however, neither of these approaches provides a coherent operational meaning in all decision problems, the way probability does.

Belief functions are closely related to the concept of upper and lower probabilities (Dempster, 1967), whereby the unique probability of an event is replaced by an upper and a lower probability. However, has always seemed to me to be a step in the wrong direction to say that because it is too difficult to specify a precise number for the probability of some event, we will specify *two* precise numbers.

There is little doubt that all of these approaches can contribute to the insights that can be gained from a thorough analysis of a particular situation. But, unfortunately there is a tendency on the part of people, including scientists, to view the world as a dichotomy comprising, on the one hand, the group to which they belong, and on the other, everyone else. Thus, those who follow a Bayesian approach consider the world to be divided into Bayesians and non-Bayesians. I suppose that

those who work with belief functions consider the world to be divided into believers and nonbelievers. It is a tribute to the talent and charisma of Professor Zadeh that so many scientists identify with a group that can only be called "fuzzy thinkers," when the rest of the world must be "clear thinkers." It is from such a dichotomous outlook that the Bayesian approach is adopted here.

In this paper we will restrict ourselves to problems in which you must determine your subjective probability of some event R , such as the probability that it will rain tomorrow in some particular location, or the probability that a particular patient has a certain disorder. It is assumed that you can consult an expert or an expert system to guide your evaluation of this probability. Thus, you will want to combine the expert's prediction, i.e., the expert's probability of R , with your own information to get your posterior probability of R .

In Section 2 we consider the question of who should be considered an expert and define the class of well-calibrated experts. In Section 3 we describe a perfect forecaster and a useless forecaster and introduce the problem of comparing well-calibrated experts. In Section 4 we induce a partial ordering in the class of well-calibrated experts by means of the concept of refinement and present several theorems that give equivalent ways of describing this partial ordering. In Section 5, we show how combining the predictions of two relatively imprecise well-calibrated experts can yield perfect predictions. Finally, in Section 6, we show how a total ordering in the class of well-calibrated experts that preserves the partial ordering

already obtained can be induced by means of the concept of strictly proper scoring rules.

2 Who is an expert?

We continue to consider the situation in which you must determine your subjective probability of some specific future event, and you can consult an expert (or an expert system) and obtain the prediction, i.e., the probability, of that expert. The question arises in this context as to just who should be regarded as an expert. Somewhat surprisingly, most articles regarding the evaluation, comparison, or combination of expert opinion, including my own articles, do not consider this question at all. Some exceptions to this silence are Morris (1974), who states that "We shall refer to ... a person who provides a judgment concerning uncertain matters as an expert," and Morris (1977), who defines an expert "to mean anyone with special knowledge about an uncertain quantity or event." Schervish (1984) writes, "... we understand the word *expert* in a very loose sense. We will assume X is an unknown quantity of interest, and we will call an expert anyone who is willing and able to state some aspect of their subjective distribution for X ." Winkler (1986) describes a "notion of goodness" of a probability appraiser which he calls "expertise" and which "relates to the degree to which the probability appraiser can approach perfect forecasts." This concept of expertise is closely related to the

concepts of calibration and refinement to be discussed in the subsequent sections of this paper.

Two extreme definitions of an expert seem possible. At one extreme, in the spirit of the authors just mentioned, we could define an expert to be anyone or any system that will give you a prediction.

At the other extreme, in this paper we will define an expert to be someone whose prediction you will simply adopt as your own posterior probability without modification. This will be the case if you believe that the expert has all of the information that you have that may be relevant to the occurrence or nonoccurrence of the event, and possibly additional information as well, and you believe that the expert processes all of this information in the way that you would process it if you had the information and the proper technical training. Of course, one way to be certain that the expert or expert system has all of the information that you have is to tell it everything that you know that is relevant.

This definition seems satisfactory if you are dealing with just a single adviser, but it raises conceptual difficulties if two advisers are present. You might very well be willing to adopt the prediction of either adviser as your own posterior probability if that was the only prediction available to you. However, after you have learned the prediction of the first adviser, you may no longer regard the second adviser as an expert according to this definition because, rather than simply accepting the second adviser's prediction, you would typically want to combine it

with the first adviser's prediction in some way to develop your own overall posterior probability. Nevertheless, in the presence of just a single advisory system, we can say in accordance with this definition that you have succeeded in building an expert system *for yourself* if you will accept its prediction in each case that it might handle.

In some of the literature (see, e.g., DeGroot and Eriksson, 1985) an expert or an expert system is said to be *well calibrated* if you will adopt its prediction as your own posterior probability. Based on the discussion that has just been given here, it would be unnecessary to use the term "well calibrated" in this paper because that property is now simply the defining characteristic of an expert. Nevertheless, for the slight cost of being redundant and the great gain of being clear about the relationship of this paper to other work on the same subject, we will use the term "well-calibrated expert" to denote an expert or a system of this type.

3 Comparing well-calibrated experts

Well-calibrated experts can exhibit a wide variety of different types of predictive behavior. Let X denote the prediction that a particular well-calibrated expert will make in a given situation. In other words, X is the probability that the expert will state for the occurrence of the event R being predicted. Before you learn the prediction of the expert, X is a random variable since you are not certain what

the expert's prediction will be.

At one extreme in the class of well-calibrated experts is the perfect forecaster who makes only the predictions $X = 0$ and $X = 1$ and who you know is always correct. In other words, this expert simply states with certainty, and without error, whether or not the event R will occur. Suppose that your prior probability of R is μ and let μ' denote your posterior probability of R after learning this expert's prediction. Then μ' will be either 0 or 1. Since $E(\mu') = \mu$, where the expectation is taken with respect to your prior distribution for μ' , it follows that you must assign probability μ to the possibility that the expert's prediction will be $X = 1$, and probability $1 - \mu$ to the possibility that $X = 0$.

At the opposite extreme in the class of well-calibrated experts is the useless forecaster whose prediction you know will be $X = \mu$. In other words, you know that this expert is simply going to repeat your own prior probability back to you. This situation arises when you regard yourself as your own expert or your own expert system.

The basic question that we will now discuss is how to compare other well-calibrated experts whose predictive behavior lies somewhere between the two extremes that have just been described. Much of the discussion to be presented is based on the material in DeGroot and Fienberg (1982, 1983, 1986) and DeGroot and Eriksson (1985), and further details, proofs, and derivations of the results can be found in those references.

In the approach to be followed here, each well-calibrated expert is characterized by your probability distribution for the expert's prediction X . For simplicity, we will assume that X is restricted to lie in a given finite subset \mathcal{X} of the closed unit interval $[0, 1]$. In effect, we are assuming that the expert's probability of R is always stated to just a fixed number of decimal places. As one example, we are all familiar with the fact that weather forecasters on American television always state their probability of precipitation to just a single decimal place. Hence, each expert can be characterized by the discrete probability function (p.f.) $\nu(x)$ of his or her prediction X .

If the expert reports $X = x$, then your posterior probability of R will be x . Hence, if your prior probability of R is μ , then as we have already indicated, $E(X) = \mu$. Thus, the comparison of all well-calibrated experts reduces to the comparison of all probability distributions on the set \mathcal{X} with mean μ .

Intuitively, it should be clear that the best experts are those about whose predictions you are most uncertain; i.e., whose predictions are most variable. If you are fairly certain in advance what prediction the expert will make — i.e., if the p.f. $\nu(x)$ is tightly concentrated around its mean μ — then there is little gain in consulting the expert. In the next section we shall make this notion rigorous.

4 Refinement

One well-calibrated expert A is said to be *at least as refined as* another well-calibrated expert B if we can simulate expert B 's prediction from expert A 's prediction and an auxiliary randomization. That is, we can simulate B 's prediction by passing A 's prediction through a noisy channel. Note that this does not mean that we can reproduce B 's actual prediction from knowing A 's prediction, but rather that we can generate a prediction that has the same stochastic properties as B 's prediction. The technical definition of this concept is based on the following notion of stochastic transformations:

A *stochastic transformation* $h(y | x)$ is a nonnegative function defined on $\mathcal{X} \times \mathcal{X}$ such that

$$\sum_{y \in \mathcal{X}} h(y | x) = 1 \quad \text{for every } x \in \mathcal{X} . \quad (4.1)$$

If the experts A and B are characterized by the p.f.'s $\nu_A(x)$ and $\nu_B(x)$, then A is defined to be *at least as refined as* B if there exists a stochastic transformation $h(y | x)$ such that

$$\sum_{x \in \mathcal{X}} h(y | x) \nu_A(x) = \nu_B(y) \quad \text{for } y \in \mathcal{X} , \quad (4.2)$$

$$\sum_{x \in \mathcal{X}} h(y | x) x \nu_A(x) = y \nu_B(y) \quad \text{for } y \in \mathcal{X} . \quad (4.3)$$

The comparison of experts in terms of the concept of refinement is very strong. In fact it can be shown that if A is at least as refined as B and you are given a

choice between learning the prediction of A or the prediction of B , you will prefer to learn that of A , regardless of the decision problem in which the prediction will be used; i.e., regardless of your utility function. The price that must be paid for using this strong method of comparison is that not all experts will be comparable. In other words, the concept of refinement introduces only a partial ordering in the class of p.f.'s $\nu(x)$ with mean μ .

It is easy to verify that the perfect forecaster described in Section 2 is at least as refined as any other well-calibrated expert, and that every well-calibrated expert is at least as refined as the useless forecaster described in that section.

We shall now describe several conditions that are equivalent to the proposition that A is at least as refined as B . Each of these equivalent conditions makes it possible to determine whether or not A is at least as refined as B without having to attempt to construct a stochastic transformation h that satisfies the definition (4.2) and (4.3).

The theory of refinement is essentially a reformulation of the theory of the comparison of statistical experiments as developed by Blackwell (1951, 1953), and from that development we can obtain further characterizations of the desired type. For any well-calibrated expert, let F denote the distribution function (d.f.) corresponding to the p.f. ν ; i.e., let

$$F(t) = \sum_{\{x \in \mathcal{X}, x \leq t\}} \nu(x) \quad \text{for } 0 \leq t \leq 1. \quad (4.4)$$

Now consider two arbitrary well-calibrated experts A and B , and let F_A and F_B denote their d.f.'s. The following result is analogous to Theorem 12.4.1 in Blackwell and Girshick (1954).

Theorem 1. Expert A is at least as refined as expert B if and only if

$$\int_0^t F_A(x) dx \leq \int_0^t F_B(x) dx \quad (4.5)$$

for all values of t in the interval $0 < t < 1$.

The relationship (4.5) between the d.f.'s F_A and F_B is known as *second-degree stochastic dominance* (see, e.g., Fishburn and Vickson, 1978).

Now let $x_0 < x_1 < \dots < x_k$ denote the finite number of points in the set X .

The following equivalent condition can be derived from (4.5):

Theorem 2. Expert A is at least as refined as expert B if and only if

$$\sum_{i=0}^{j-1} (x_j - x_i) [\nu_A(x_i) - \nu_B(x_i)] \geq 0 \quad \text{for } j = 1, \dots, k-1 \quad (4.6)$$

Another equivalent condition can be presented in terms of the Lorenz curve, which is defined as follows (see, e.g., Gastwirth, 1971):

Suppose that F is the d.f. of an arbitrary non-negative random variable and, for $0 < \mu < 1$, define

$$F^{-1}(\mu) = \inf\{t : F(t) \geq \mu\} \quad (4.7)$$

The function F^{-1} is called the *quantile function* corresponding to the d.f. F . If μ again represents the mean of the distribution with d.f. F , then the Lorenz curve

$L(t)$ corresponding to the d.f. F is given by

$$L(t) = \frac{1}{\mu} \int_0^t F^{-1}(u) du \quad \text{for } 0 \leq t \leq 1 \quad (4.8)$$

For any d.f. F , the Lorenz curve $L(t)$ is a convex, nondecreasing function on the interval $0 \leq t \leq 1$ such that $L(0) = 0$ and $L(1) = 1$. When F is the d.f. of a discrete distribution concentrated on just a finite number of points, as is true of all the d.f.'s we are considering in this paper, then $L(t)$ is also piecewise linear.

Now consider again two well-calibrated experts A and B , and let L_A and L_B denote the Lorenz curves corresponding to their d.f.'s F_A and F_B .

Theorem 3. Expert A is at least as refined as expert B if and only if

$$L_A(t) \leq L_B(t) \quad \text{for all } 0 \leq t \leq 1 \quad (4.9)$$

The next two equivalent conditions that will be presented give additional insight into the relationship of refinement, but do not provide a direct way of verifying that this relationship holds.

Theorem 4. Expert A is at least as refined as expert B if and only if

$$\sum_{x \in \mathcal{X}} g(x) \nu_A(x) \geq \sum_{x \in \mathcal{X}} g(x) \nu_B(x) \quad (4.10)$$

for every convex function g on the interval $[0, 1]$.

Theorem 5. Expert A is at least as refined as expert B if and only if there exists a stochastic transformation $\eta(x | y)$ such that

$$\sum_{y \in \mathcal{Y}} \eta(x | y) \nu_B(y) = \nu_A(x) \quad \text{for } x \in \mathcal{X} \quad (4.11)$$

$$\sum_{x \in \mathcal{X}} x \eta(x | y) = y \quad \text{for } y \in \mathcal{Y} . \quad (4.12)$$

Theorem 5 is interesting because it shows that although the definition of A being more refined than B depends on the existence of a stochastic transformation from x to y satisfying certain properties, there is an equivalent condition in terms of a stochastic transformation from y to x satisfying certain other properties.

Results of the type that have been presented here are closely related to the theory of majorization, as described, for example by Marshall and Olkin (1979). Indeed, one final equivalent way of saying that A is at least as refined as B is to say that the p.f. ν_A majorizes the p.f. ν_B .

5 Two experts

As we have stated, if expert A is at least as refined as expert B and you are given a choice between learning either the prediction of expert A or the prediction of expert B (at the same cost), then you will always prefer to learn that of A , regardless of the use you are going to make of the prediction. However, it should also be emphasized that if you can learn the prediction of expert B in addition to the opinion of expert A , then that additional information will often be useful in the sense that it will further modify your posterior probability of R . This is possible because the relationship that A is at least as refined as B depends only on the marginal p.f.'s ν_A and ν_B of each expert. When we consider the joint p.f.

of their predictions, and the conditional probability of R given both predictions.

the situation can change drastically, as the following simple example shows.

Let X and Y denote the predictions of experts A and B , respectively, and suppose that both X and Y can have only the two possible values $\frac{1}{4}$ and $\frac{3}{4}$.

Suppose also that the joint distribution of X and Y is as follows:

$$\begin{aligned} Pr\left(X = \frac{1}{4}, Y = \frac{1}{4}\right) &= \frac{1}{16} \\ Pr\left(X = \frac{1}{4}, Y = \frac{3}{4}\right) &= Pr\left(X = \frac{3}{4}, Y = \frac{1}{4}\right) = \frac{3}{16} \\ Pr\left(X = \frac{3}{4}, Y = \frac{3}{4}\right) &= \frac{9}{16} \end{aligned} \quad (5.1)$$

and that

$$\begin{aligned} Pr\left(R \mid X = \frac{1}{4}, Y = \frac{1}{4}\right) &= Pr\left(R \mid X = \frac{3}{4}, Y = \frac{3}{4}\right) = 1 \\ Pr\left(R \mid X = \frac{1}{4}, Y = \frac{3}{4}\right) &= Pr\left(R \mid X = \frac{3}{4}, Y = \frac{1}{4}\right) = 0 \end{aligned} \quad (5.2)$$

We will now show both expert A and expert B are well calibrated:

$$\begin{aligned} Pr\left(R \mid X = \frac{1}{4}\right) &= Pr\left(R \mid X = \frac{1}{4}, Y = \frac{1}{4}\right) Pr\left(Y = \frac{1}{4} \mid X = \frac{1}{4}\right) \\ &\quad + Pr\left(R \mid X = \frac{1}{4}, Y = \frac{3}{4}\right) Pr\left(Y = \frac{3}{4} \mid X = \frac{1}{4}\right) \\ &= \frac{1}{16} / \left(\frac{1}{16} + \frac{3}{16}\right) = \frac{1}{4} \end{aligned} \quad (5.3)$$

and

$$\begin{aligned}
Pr\left(R \mid X = \frac{3}{4}\right) &= Pr\left(R \mid X = \frac{3}{4}, Y = \frac{1}{4}\right) Pr\left(Y = \frac{1}{4} \mid X = \frac{3}{4}\right) \\
&+ Pr\left(R \mid X = \frac{3}{4}, Y = \frac{3}{4}\right) Pr\left(Y = \frac{3}{4} \mid X = \frac{3}{4}\right) \\
&= \frac{9}{16} / \left(\frac{3}{16} + \frac{9}{16}\right) = \frac{3}{4} .
\end{aligned} \tag{5.4}$$

Together, (5.3) and (5.4) show that expert A is well calibrated since the posterior probability of R given A 's prediction $X = x$ is simply x itself. The analogous calculation shows that

$$Pr\left(R \mid Y = \frac{1}{4}\right) = \frac{1}{4} \quad \text{and} \quad Pr\left(R \mid Y = \frac{3}{4}\right) = \frac{3}{4} . \tag{5.5}$$

which proves that expert B is also well calibrated.

Hence, if you learn either the prediction of expert A or of expert B , but not both, that prediction will become your posterior probability of R . On the other hand, if you could learn the predictions of both experts A and B , then (5.2) reveals that you would be certain whether or not R will occur. In summary, in this example the combination of two relatively imprecise well-calibrated experts can be completely informative as to whether R will occur.

6 Ordering all experts

As we have discussed, the relation of refinement induces only a partial ordering of the class of well-calibrated experts. It is natural to try to obtain a total ordering of

this class, and one way to accomplish this ordering is to assign a numerical measure of quality to the experts. Thus, we wish to assign a value $m(\nu)$ to each p.f. ν defined on the set \mathcal{X} and having mean μ . The values $m(\nu)$ should be assigned in such a way that the "better" experts receive the larger values. We interpret this requirement to mean that if expert A is at least as refined as expert B , then $m(\nu_A) \geq m(\nu_B)$, with strict inequality unless the p.f.'s ν_A and ν_B are identical. (A function m with this property is called Schur-convex; see, for example, Marshall and Olkin, 1979, or DeGroot and Eriksson, 1985).

One way to develop appropriate measures of quality is to invoke the concept of *strictly proper scoring rules* (see, e.g., Staël von Holstein, 1970; Savage, 1971; Winkler, 1977 and 1986; or DeGroot and Fienberg, 1983). Suppose that if an expert's prediction is x and the event R actually occurs, the expert will receive a score $g_1(x)$; whereas if R does not occur, the expert will receive a score $g_2(x)$. We assume that the expert desires to maximize his or her score, so we will assume that $g_1(x)$ is an increasing function of x and that $g_2(x)$ is a decreasing function of x . Together, the pair of functions (g_1, g_2) is said to form a *scoring rule*.

Consider now the possibility that although an expert's actual subjective probability of R is p , the prediction that the expert reports is x , where x is not necessarily equal to p . (This possibility clearly exists for a human expert, although it may not exist if the expert is actually an expert system, i.e., a computer program.)

Under these conditions, the expert's expected score is

$$pg_1(x) + (1 - p)g_2(x) \quad . \quad (6.1)$$

The scoring rule (g_1, g_2) is said to be *strictly proper* if $x = p$ is the unique value of x that maximizes (6.1).

The idea behind strictly proper scoring rules is that they are supposed to encourage the expert to report an "honest" prediction because only such a report maximizes the expert's expected score. Of course, for this idea to be effective, one must somehow motivate the expert to want to maximize his or her expected score. Nevertheless, strictly proper scoring rules are precisely the appropriate class of scoring rules that should be considered in order to obtain measures of quality m having the property that we desire.

Suppose therefore that (g_1, g_2) is a strictly proper scoring rule, and let

$$g(x) = xg_1(x) + (1 - x)g_2(x) \quad . \quad (6.2)$$

Then it can be shown (Savage, 1971) that $g(x)$ must be a strictly convex function on the interval $0 \leq x \leq 1$. Now let the measure of quality m be defined for any p.f. ν by the relation

$$m(\nu) = \sum_{x \in \mathcal{X}} g(x)\nu(x) \quad . \quad (6.3)$$

In other words, the measure of quality $m(\nu)$ that you assign to a well-calibrated expert who is characterized by the p.f. ν is simply your expectation of the score

that the expert will receive, before you learn the expert's prediction X . The next result now follows from Theorem 4 and the extra consideration that g is not only convex, but strictly convex.

Theorem 6. If expert A is at least as refined as expert B , then $m(\nu_A) \geq m(\nu_B)$, with strict inequality unless $\nu_A(x) = \nu_B(x)$ for all $x \in \mathcal{X}$.

In summary, each choice of a strictly proper scoring rule leads to a (strictly) Schur-convex measure of quality m , by means of the construction (6.2) and (6.3).

The two most widely known strictly proper scoring rules for the evaluation of forecasters are the Brier scoring rule (Brier, 1950), defined by the relations

$$g_1(x) = -(x - 1)^2, \quad g_2(x) = -x^2, \quad (6.4)$$

and the logarithmic scoring rule (Good, 1952), defined by the relations

$$g_1(x) = \log x, \quad g_2(x) = \log(1 - x). \quad (6.5)$$

Others are described in the references already cited in this paper.

REFERENCES

Blackwell, D. (1951). Comparison of experiments. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 93-102. University of California Press, Berkeley.

Blackwell, D. (1953). Equivalent comparison of experiments. *Annals of Mathematical Statistics* **24**, 265-272.

Blackwell, D. and Girshick, M. A. (1954). *Theory of Games and Statistical Decisions*. John Wiley, New York.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1-3.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

DeGroot, M. H. and Eriksson, E. A. (1985). Probability forecasting, stochastic dominance, and the Lorenz curve. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.), 99-118. North-Holland, Amsterdam.

DeGroot, M. H. and Fienberg, S. E. (1982). Assessing probability assessors: calibration and refinement. In *Statistical Decision Theory and Related Topics III*, Vol. 1 (S. S. Gupta and J. O. Berger, eds.), 291-314. Academic Press, New York.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation

of forecasters. *The Statistician* **32**, 12-22.

DeGroot, M. H. and Fienberg, S. E. (1986). Comparing probability forecasters: basic binary concepts and multivariate extensions. In *Bayesian Inference and Decision Techniques* (P. K. Goel and A. Zellner, eds.) 247-264. Elsevier Science Publishers (North-Holland), New York.

Dempster, A. P. (1967). Upper and lower probabilities induced by multivalued mapping. *Annals of Mathematical Statistics* **38**, 325-339.

Fishburn, P. C. and Vickson, R. G. (1978). Theoretical foundations of stochastic dominance. In *Stochastic Dominance* (G. A. Whitmore and M. C. Findlay, eds.), 39-113. Lexington Books, Lexington, Massachusetts.

Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica* **39**, 1037-1039.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B* **14**, 107-114.

Lindley, D. V. (1987). The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. *Statistical Science*, **3**, No. 1, to appear.

Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York.

Morris, P. A. (1974). Decision analysis expert use. *Management Science*, **20**, 1233-1241.

Morris, P. A. (1977). Combining expert judgments: a Bayesian approach. *Management Science*, **23**, 679-693.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley, New York.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, **66**, 783-801.

Schervish, M. J. (1984). Calibration and combining expert judgments. Technical Report No. 294, Department of Statistics, Carnegie Mellon University.

Shafer, G. (1976). *Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey.

Shafer, G. (1982). Belief functions and parametric models. *Journal of the Royal Statistical Society. Series B*, **44**, 322-352 (with discussion).

Shafer, G. (1987). Probability judgment in artificial intelligence and expert systems. *Statistical Science*, **2**, No. 1 to appear.

Stael von Holstein, C.-A. S. (1970). *Assessment and Evaluation of Subjective Probability Distributions*. Economic Research Institute, Stockholm School of Economics, Stockholm.

Winkler, R. L. (1977). Rewarding expertise in probability assessment. In *Decision Making and Change in Human Affairs* (H. Jungermann and G. de Zeeuw, eds), 127-140, D. Reidel, Dordrecht, Holland.

Winkler, R. L. (1986). On "good probability appraisers." In *Bayesian Inference and Decision Techniques* (P. K. Goel and A. Zellner, eds.), 265-278. Elsevier

Science Publishers (North-Holland). New York.

Zadeh, L. A. (1979). Fuzzy sets and information granularity. *Advances in Fuzzy Set Theory and Applications* (M. M. Gupta, R. K. Ragade, and R. R. Yager, eds.) North-Holland, Amsterdam.

Zadeh, L. A. (1983). The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Systems II*, 199-227. 